Profiling Structured Product Labeling with NDF-RT and RxNorm

Qian Zhu[1,*] Guoqian Jiang[1] and Christopher G Chute[1]

[1]Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

ABSTRACT

The Structured Product Labeling (SPL) is a document markup standard approved by Health Level Seven (HL7) and adopted by United States Food and Drug Administration (FDA) as a mechanism for exchanging drug product information. The SPL includes rich information about FDA approved clinical drugs, however, lack of linkage to standard drug ontologies hinders its meaningful use. In this paper, we present a framework that intends to map SPL drug labels with existing drug ontologies, NDF-RT (National Drug File Reference Terminology) and RxNorm. We also use existing categorical annotations to classify SPL drug labels into corresponding classes. We established the classification and relevant linkage for SPL drug labels using the following 3 approaches. First, we retrieved NDF-RT categorical information from the External Pharmacologic Class (EPC) indexing SPLs. Second, we used the RxNorm and NDF-RT mappings to classify and link SPLs with NDF-RT categories. Third, we profiled SPLs using RxNorm term type information. In the implementation process, we employed a Semantic Web technology framework, in which we stored the data sets from RxNorm, NDF-RT and SPLs into a RDF triple store, and executed SPARQL queries to retrieve data from customized SPARQL endpoints.

1 INTRODUCTION

Structured Product Labeling (SPL) encodes rich clinical drug knowledge, however, it is not easy to integrate the SPL labels with other data sources due to proprietary formats, and different conceptual models. This is a common scenario occurring in the biomedical domain, where dozens of public resources involve laborious processes to manually annotate data. This is mostly because they are using heterogeneous code systems to represent their data. Hence, data normalization and building all possible linkages among these data sets will make data interoperation and integration feasible.

The objective of the present study is to map SPL drug labels using two major standard drug ontologies: the Veterans Administration's (VA) NDF-RT[1] and the National Library of Medicine's (NLM) RxNorm[2]. Our investigation was guided by answering the following research questions: (1) how SPL drug labels are covered and connected by RxNorm and NDF-RT; (2) how to utilize RxNorm/NDF-RT drug resources to map SPL drug labels from the drug class perspective; (3) how to leverage Semantic Web technology to accomplish the implementation task.

The paper is organized into the following sections. First, we introduce background information for SPL, NDF-RT, RxNorm and Semantic Web technology in the Background section; Second, in the Methods section, we introduce three main parallel approaches on SPL drug label profiling; Third, we illustrate our results generated from each step in the Results section, and then followed by Discussion and Conclusion.

2 BACKGROUND

2.1 Structured Product Labeling (SPL)

The Structured Product Labeling (SPL)[3] is a document markup standard approved by Health Level Seven (HL7)[4] and adopted by FDA as a mechanism for exchanging product information. SPL defines the human readable label documents that contain structured content of labeling (all text, tables and figures) for a product, along with additional machine readable information (i.e., drug listing data elements including information about the product and the packaging). SPL labels used in this study were extracted from NLM DailyMed website[5].

2.2 National Drug File Reference Terminology (NDF-RT)

NDF-RT[1] is used for modeling drug characteristics including ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases.

In support of SPL initiative, a non-hierarchical collection of External Pharmacologic Class (EPC) concepts has been added to NDF-RT in parallel and analogous with the VA Drug Classification hierarchy. These concepts are distinguished by an "[EPC]" tag suffixed to their preferred names. Role relationships originating from these EPC concepts target concepts from the NDF-RT MoA, PE, and CI hierarchies that are selected by the FDA to index their EPC for SPL purposes[6].

* To whom correspondence should be addressed: zhu.qian@mayo.edu

[1] NDF-RT: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/

[2] RxNorm: http://www.nlm.nih.gov/research/umls/rxnorm/

[3] SPL: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm

[4] HL7: http://www.hl7.org/

[5] DailyMed: http://dailymed.nlm.nih.gov

[6] NDF-RT documentation: evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf

2.3 RxNorm

RxNorm[2] provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used. Also it makes links to a number of vocabularies, such as SPL, NDF-RT, MeSH, and etc. The "SAB" code is defined by RxNorm to differentiate the different sources aggregated into RxNorm. For example, "SAB=MTHSPL" indicates that the corresponding concept is absorbed from SPL and "SAB=NDFRT" indicating the source from NDF-RT. These two sources were used in this study. RxNorm defines term type "TTY" to indicate the role an atom plays in its source. The term types are assigned based on source documentation or NLM understanding of the source. Table 1 shows a list of term types "TTYs" used in this study with their names and descriptions.

TTY	Name	Description
SBD	Semantic Branded Drug	Ingredient + Strength + Dose Form + Brand Name
SCD	Semantic Clinical Drug	Ingredient + Strength + Dose Form
IN	Ingredients	A compound or moiety that gives the drug its distinctive clinical properties.
PIN	Precise Ingredient	A specified form of the ingredient that may or may not be clinically active.
BPCK	Brand Name Pack	{# (Ingredient Strength Dose Form) / # (Ingredient Strength Dose Form)} Pack [Brand Name]
GPCK	Generic Pack	{# (Ingredient + Strength + Dose Form) / # (Ingredient + Strength + Dose Form)} Pack
BN	Brand Name	A proprietary name for a family of products containing a specific active ingredient.
MIN	Multiple Ingredients	Two or more ingredients appearing together in a single drug preparation, created from SCDF.
SY	Synonym	Synonym of another TTY, given for clarity.
TMSY	Tall Man Lettering Synonym	Tall Man Lettering synonym of another TTY, given to distinguish between commonly confused drugs.

Table 1. A list of RxNorm term types "TTYs" with names and descriptions (Source from RxNorm Documentation[7])

In this study, we used the following two files that were downloaded from RxNorm website in April 9, 2012. 1) RXNCONSO.RRF. The file includes all connections with source vocabularies. 2) RXNSAT.RRF. The file includes all source vocabulary attributes that do not fit into other categories to search for the information about drug categories and connections among RxNorm, NDF-RT and SPL.

2.4 Semantic Web Technology

The Resource Description Framework (RDF) [8], a W3C recommendation, is a directed, labeled graph data format for representing information in the Web. SPARQL is a query language for RDF graphs (Clement et al. 2009). Triple store is a database for the storage and retrieval of RDF metadata, ideally through standard SPARQL query language. The Web Ontology Language (OWL) is a standard ontology language for the Semantic Web[9]. Semantic Web technology supports flexible, extensible and evolvable knowledge transfer and reuse. It has been widely used in biomedical domains to formalize and model medical and biological systems (Neumann et al. 2004; O'Connor et al. 2007b; David et al. 2011). In the present study, we adopted it as the core technology in our implementation step. The SPL data used in this study is stored in a RDF triple store and by executing SPARQL queries to retrieve the desirable information.

3 METHODS

3.1 System Architecture

There are four primary modules in the system, comprising of 1) data transformation module; 2) data persistence module; 3) SPL profiling module, in which SPL drug labels are profiled by EPC, NDF-RT and RxNorm; 4) standardized drug/drug class network module. Figure 1 shows system architecture of the four modules.

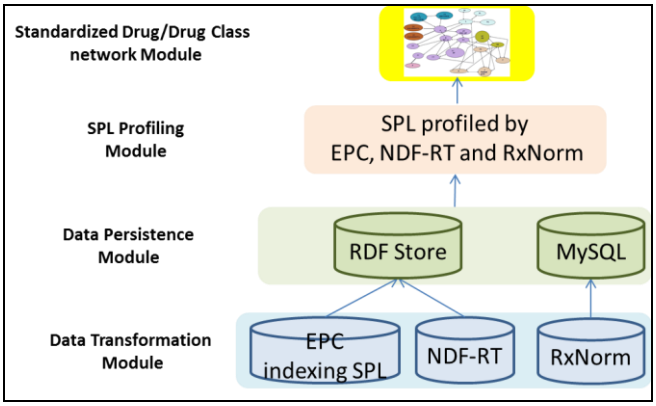

Fig.1. A diagram illustrating our system architecture.

For the data transformation module, data reformatting steps were performed for SPL, NDF-RT and RxNorm individually prior to loading the data into a RDF tripe store and a MySQL database since they have different data formats - original SPLs in XML format, NDF-RT available in OWL (Antoniou et al. 2004), and RxNorm in the UMLS Rich Representation Format (RRF). A XML2RDF sub module[10] takes the input rendered in the XML format, and outputs the result in the RDF format through a transparent transformation service. NDF-RT in OWL was loaded into RDF store directly. Although the RxNorm data is available in RRF files, RxNorm provides a MySQL script to help load the data into a MySQL database. In this study, we queried against from a MySQL database for RxNorm data.

For the persistence module, we implemented an open source RDF store "4Store" that is developed by Garlik[11] to host the SPL and NDF-RT data. After loading RDF triples into the RDF store, we implemented a SPARQL endpoint that

[7] RxNorm Documentation: http://www.nlm.nih.gov/research/umls/rxnorm/overview.html#

[8] The RDF: http://www.w3.org/RDF/; last visited April 12, 2012.

[9] OWL: http://www.w3.org/TR/owl-features/

[10] XML2RDF: rhizomik.net/redefer/xml2rdf/

[11] 4Store: http://4store.org/

provides standard SPARQL query service against the RDF store.

For the drug/drug class network module, we incorporated SPL profiling results with our previous standardized drug work [Zhu et al. 2012]. We will integrate more drug / drug class resources, like PharmGKB[12], DrugBank[13], or National Drug Code (NDC) [14] into this network, which will be visualized by Cytoscape[15], a general platform for complex network analysis and visualization.

3.2 Profiling by EPC Classes

The EPC indexing SPLs were stored in a RDF triple store. We executed SPARQL query against the triple store to extract the setId (unique identifier of the SPL), NUI (unique identifier of the NDF-RT), and the relevant role relationships describing and defining concepts according to their relationships with other concepts, mapped by NDF-RT from each SPL. Figure 2 shows the SPARQL query used to extract the EPC indexing SPL and EPC classes with related NDF-RT concepts associated with their role relationships.

```
PREFIX epc:<urn:hl7-org:v3#>
SELECT ?setId ?nui ?displayName
   {GRAPH <http://fda.gov/spl/epc>
   {   ?xml epc:setId ?setIdSection .
       ?setIdSection epc:root ?setId .
       ?xml epc:component ?component .
       ?component epc:structuredBody ?structuredBody .
       ?structuredBody epc:component ?subcomponent .
       ?subcomponent epc:section ?section .
       ?section epc:subject ?subject .
       ?subject epc:identifiedSubstance ?sub .
       ?sub epc:identifiedSubstance ?sub1 .
       ?sub1 epc:asSpecializedKind ?kind .
       ?kind epc:generalizedMaterialKind ?kind1 .
       ?kind1 epc:code ?code .
       ?code epc:displayName ?displayName .
      FILTER (regex(?displayName, \"EPC\", \"i\") ||
regex(?displayName, \"PE\", \"i\") || regex(?displayName, \"Chemical\",
\"i\"))
       ?code epc:code ?nui .
      FILTER regex(?nui, \"N\", \"i\")
}}
```

Fig.2. SPARQL query to extract EPC classes and NDF-RT concepts with their role relationships

The NDF-RT concepts whose names replicate each EPC string are mapped via role relationships to appropriate MoA, PE, or Chemical/Ingredient concepts in NDF-RT. An EPC indexing SPL label is possibly corresponding to multiple NDF-RT EPC classes, and for each EPC class, multiple role relationships might be mapped to.

3.3 Profiling by RxNorm and NDF-RT mapping

RxNorm reflects and preserves the meanings, concept names, and relationships from different copyright holders. Hence, the objective of this step was to use existing annotations in RxNorm to categorize SPL drug labels, and to make connections between SPLs and RxNorm/NDF-RT. As RxNorm (RXNCONSO and RXNSAT) were pre-loaded in a MySQL database, we executed the SQL clauses to extract the data. We used two RxNorm-integrated resources, SPL and NDF-RT in this step.

For the source SPL (i.e. "SAB = MTHSPL"), we executed SQL queries to extract the SPL concepts along with "TTY" and RxCUI (RxNorm unique identifier) from RXNCONSO table. In order to connect RxNorm with SPL labels, we queried RXNSAT table to retrieve corresponding SPL setIds for each RxCUI.

For the source NDFRT, (i.e., "SAB = NDFRT"), SQL queries were executed to extract the NDF-RT concepts along with NUI and preferred name. Each preferred name includes additional "KIND" information from NDF-RT. For example, an entry with [RxCUI = "4278"] corresponds to the [NUI = "N0000006373"] with preferred name "Famotidine [Chemical/Ingredient]". We grouped the entries into NDF-RT kind "Chemical/Ingredient".

It is worthy to note that RxCUIs are available in "RXCUI" column from both RXNCONSO and RXNSAT tables. NUI is stored in "SCUI (Source asserted concept identifier)" column, while no setId available in SCUI column. Hence, to establish the linkage between SPL and RxNorm, we searched RXNSAT to retrieve a list of setIds for each given RxCUI.

4 RESULTS

Total number of SPLs from DailyMed is 36,568. 1,247 EPC indexing SPLs were downloaded as of April 12, 2012 and loaded into RDF triple store. Meanwhile, we have RxNorm snapshot downloaded at the same time. "RXNCONSO.RRF" with 965,968 concepts and "RXNSAT.RRF" with 6,221,513 entries were loaded into a MySQL database.

4.1 Results from EPC Classes

Each EPC indexing SPL can be mapped to multiple NDF-RT concepts with different or same role relationships. For example, a EPC indexing SPL "MUMPS VIRUS STRAIN B LEVEL JERYL LYNN LIVE ANTIGEN" with setId "433534fd-7318-45ec-b42c-9f35e7a7140e" has EPC class "Live Mumps Virus Vaccine", which is mapped to three NDF-RT concepts by role relationships, "has_PE" with "Actively Acquired Immunity", "has_Chemical_Structure" with "Vaccines, Attenuated", and "has_Chemical_Structure" with "Mumps Vaccine".

The mapping results are listed in Table 2. The coverage shown in brackets in the Table 2 is calculated by number of concepts from NDF-RT and SPL with each role relationship divided by the total number of concepts from these two resources. Total 497 EPC classes are found in the NDF-RT

[12] PharmGKB: www.pharmgkb.org

[13] DrugBank: www.drugbank.ca

[14] NDC: http://www.fda.gov/drugs/informationondrugs/ucm142438.htm

[15] Cytoscape: http://www.cytoscape.org

RDF repository, indicating that 71.2% (354 out of 497) EPC classes have been integrated into SPL.

Category	Num. of unique NUIs	Num. of unique setIds
EPC	354(0.8%)	853(2.3%)
Chemical/ Ingredient	154(0.3%)	342(0.9 %)
PE	70(0.2%)	201(0.6%)
MoA	7(0.01%)	10(0.03%)
Total	585(1.2%)	858(2.4%)

Table 2. Mapping results by EPC classes (Category: NDF-RT categories extracted from EPC indexing SPL; NUI: identifier of NDF-RT; setId: identifier of SPL)

4.2 Results from RxNorm vs. NDF-RT mappings

We extracted 41,343 unique RxNorm entries from the source "NDFRT" and their corresponding NUIs and preferred names. We identified 6,611 unique NUIs that correspond to 35,094 unique SPL setIds retrieved from RXNSAT file, which includes the connections between RxCUI and setIds. In this step, we utilized the NDF-RT category information to profile the SPLs. Only three categories from NDF-RT (VA Product, Chemical/Ingredient, and EPC) have been included in RxNorm and NDF-RT mappings, hence, we extracted these three categories along with NUIs and setIds. The statistical results are listed in Table 3. Comparing with the 36,568 SPL labels from DailyMed and 47,075 concepts from NDF-RT, 96% SPLs have been covered by RxNorm and NDF-RT mappings, and only 14.0% NDF-RT concepts have been linked to the SPL labels.

Category	Num. of unique NUIs	Num. of unique setIds
VA Product	4,880(10.4%)	20,937(57.3%)
Chemical/Ingredient	1,730(3.7%)	34,788(95.1%)
EPC	1(0.002%)	14(0.04%)
Total	6,611(14.0%)	35,094(96.0%)

Table 3. Mapping results by RxNorm/NDF-RT mappings (Category: NDF-RT categories from NDF-RT/RxNorm mappings; NUI: identifier of NDF-RT; setId: identifier of SPL)

4.3 Results from RxNorm

We first identified 35,480 unique SPL entries from RxNorm using the source "SPL". And then we identified 15,615 unique RxCUIs that correspond to the 35,480 unique SPL setIds by searching for RXNSAT table. We used the term types "TTYs" to classify SPL labels into ten categories. Each category with the number of unique RxCUIs and unique setIds and their coverage has been listed in the Table 4. Comparing with the 36,568 SPL labels from DailyMed and 965,968 concepts from RxNorm, 97.0% SPLs have been covered by RxNorm and NDF-RT mappings, and only 1.6% RxNorm codes have been linked to the SPL labels. It is worthy to note that SY and TMSY denote synonyms of another TTY, so there are overlaps between SY/TMSY and other TTYs.

TTY	Num. of unique RxCUIs	Num. of unique setIds
SBD	6714(0.70%)	7087(19.38%)
SCD	5981(0.6%)	21127(21.9%)
IN	1,834(0.2%)	34,038(93.1%)
PIN	773(0.08%)	18,498(50.6%)
BPCK	261(0.03%)	259(0.7%)
GPCK	48(0.005%)	150(0.04%)
BN	2(0.0002%)	5(0.01%)
MIN	2(0.0002%)	3(0.008%)
SY*	11,489(1.2%)	21,438(58.6%)
TMSY*	1,810(0.2%)	4,208(11.5%)
Total	15,615(1.6%)	35,480(97.0%)

Table 4. Mapping results from RxNorm (TTY denotes term types from RxNorm; RxCUI denotes identifier of RxNorm; setId denotes identifier of SPL. * denotes the concepts with SY and TMSY have overlaps with the concepts in other TTYs)

4.4 Network Visualization

The SPL profiles using RxNorm and NDF-RT not only demonstrate the connections among these three resources, but also help to establish a drug/drug class network among them. Within this network, the nodes are consisted of concepts from SPLs, RxNorm or NDF-RT as target or source nodes; the edges are consisted of the category information retrieved in the above steps. We are exploring to utilize the Cytoscape[15] as a visualization tool to visualize the network.

5 DISCUSSION

Since SPL labels contain a large portion of clinical drugs, and chemical/ingredients, along with other possible drug categories, they have been demonstrated as a very useful drug knowledge resource, especially in clinical drug applications such as Adverse Drug Events (ADEs) detection from electronic medical records (EMRs). Actually a number of studies are emerging recently to semantically annotate ADEs information using the SPL labels for the purpose of drug safety surveillance. For example, in a project called SIDER, a public, computer-readable side effect resource that connects 888 drugs to 1450 side effect terms was developed using the SPL labels (Kuhn et al. 2010). In a system called ADESSA, for another example, the ADEs were extracted from the SPL labels and mapped to the MedDRA terms and concepts, then utilized the UMLS to generate mappings between the MedDRA terms and the SNOMED CT concepts (Duke et al. 2010). In a project at Mayo Clinic, the SPL labels were used in a framework for building a standardized ADE knowledge base known as *ADEpedia*[16] through combining ontology-based approaches with Semantic Web technology (Jiang et al. 2011). In addition, Schadow conducted some other studies to evaluate the impact of SPL for medication knowledge management

[16] ADEPedia: http://adepedia.org

(Schadow 2007). And Schadow also had successfully aligned SPL with associated terminologies to make drug-intolerance (allergy) decision support in computerized provider order entry (CPOE) systems in 2008 (Schadow 2008).

In this paper, we conducted the study with SPLs from drug and drug classification perspective. We have successfully mapped SPL labels with NDF-RT and RxNorm using NDF-RT drug/drug class information and clinical drug identification information from RxNorm. For the SPLs and NDF-RT mappings, 96.0% of SPL drug labels are linked with NDF-RT categories; whereas for the SPLs and RxNorm mappings, 97.0% of SPL drug labels are linked to RxNorm codes.

The SPL mappings reported in this study reveal the SPL coverage by NDF-RT and RxNorm, which can potentially facilitate the future integration among SPL, NDF-RT and RxNorm for normalizing drug information as much as possible. In addition, the SPL mappings also enable the integration of SPL drug labels into a drug and drug class network developed in our previous study (Zhu et al. 2012).

Semantic Web technology plays a key role in the implementation of our system. We represented the drug data in RDF triples and hosted the RDF triples in a RDF triple store to make the data integration, data management more feasible. In addition, running SPARQL queries against RDF store through SPARQL endpoint simplified our efforts on linking SPL drug labels to NDF-RT by using the predicates defined in the RDF triples. In the future, we will employ D2R (Bizer et al. 2003) server for the RxNorm RDF transformation through describing mappings between the RRF-based relational database schema and a RDF data model.

6 CONCLUSION

In this study, we have successfully mapped SPL drug labels with RxNorm and NDF-RT. In total, 99.4% SPL drug labels have been linked with two drug ontologies: RxNorm and NDF-RT. The coverage for individual RxNorm and NDF-RT is 97.0% and 96.0%. Since NDF-RT provides rich drug/drug class information, we were able to map SPLs to NDF-RT from drug class perspective. Meanwhile, we utilized the clinical meaningful drug term types from RxNorm to profile SPLs.

We will continue the following investigations in the future. 1) Since existing SPL drug labels have been classified into a number of categories, including over-the-counter (OTC), prescription, animal, and so on, we will explore to integrate the information into the current drug network. 2) We will build a backbone drug network based on NDF-RT and integrate the network with more drug resources, like DrugBank, NDC, PharmGKB, etc. 3) We will explore to build linkages between the drug/drug class and relevant phenotype/genotype.

ACKNOWLEDGEMENTS

This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS - U19 GM61388) and the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

REFERENCES

Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. (2010) A side effect resource to capture phenotypic effects of drugs. *Molecular System Biology*; **6**:343-9.

Duke J. D., Friedlin J. (2010) ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. *AMIA Annu Symp Proc.* **2010**:177-81.

Jiang G, Solbrig H. R, Chute C.G. (2011) ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. *AMIA Annu Symp Proc.* **2011**:607-16.

Schadow G. (2007) Assessing the impact of HL7/FDA Structured Product Label (SPL) content for medication knowledge management. *AMIA Annu Symp Proc.* 646–50.

Schadow G. (2008) Structured Product Labeling Improves Detection of Drug-intolerance Issues. *J Am Med Inform Assoc*,**16**:211-219

Neumann E.K, Miler E., Wilbanks J. (2004) What the semantic web could do for the life sciences. Drug Discovery Today, 2:228-34.

Wild D. J., Ding Y., Sheth A.P., Harland L., Gifford E.M., Lajiniess M.S. (2011) Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. *Drug Discovery Today*, http://dx.doi.org/10.1016/j.bbr.2011.03.031

Clement J., Nigam S. H., Mark M.A. (2009). The Open Biomedical Annotator. *AMIA Summit on Translational Bioinformatics*; San Francisco, CA, USA. 56–60.

Antoniou G, van Harmelen F (2004). Web Ontology Language: OWL. In: Staab S, Studer R, editors. *Handbook on ontologies.* Berlin; New York: Springer; 67-92.

Bizer, C. (2003). D2R MAP-A Database to RDF Mapping Language. *In: Proceedings of the 12th International World Wide Web Conference*. 20–24.

Schadow G. (2007) Assessing the impact of HL7/FDA Structured Product Label (SPL) content for medication knowledge management. *AMIA Annu Symp Proc.* 646–50.

O'Connor, M. J., Shankar, R. D., Tu, S. W., Parrish, D. B., Das, A. K., & Musen, M. A. (2007b). Using semantic web technologies for knowledge-driven querying of biomedical data. In: *Proc. 11th Conf. on Artificial Intelligence in Medicine (AIME2007)*, LNAI **4594**:267–276.

Zhu Q, Jiang G, Chute, C.G. (2012) A Standardized Drug and Drug Class Universal Network. Accepted by *AMIA Annu Symp Proc.2012*